Malays. Appl. Biol. (2015) 44(1): 161–165

IN SILICO CHARACTERISATION OF THE *Glaciozyma antarctica* GENOME: MINING THE MOLECULAR CHAPERONES

NUR ATHIRAH, Y.[1], ABU BAKAR, F.D.[1], ILLIAS, R.M.[2], MAHADI, N.M.[3] and MURAD, A.M.A.[1*]

[1]*School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia*
[2]*Department of Bioprocess Engineering, Faculty of Chemical and Natural Resources Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia*
[3]*Malaysia Genome Institute, Jalan Bangi, 43000 Kajang, Selangor, Malaysia*
[*]*Email: munir@ukm.edu.my*

ABSTRACT

Glaciozyma antarctica is a psychrophilic yeast isolated from the Antarctic sea ice. In this study, we performed a *de novo* characterisation of molecular chaperones from *G. antarctica* genome datasets. A total of 7857 genes that code for various types of proteins have been predicted from the *G. antarctica* genome sequence. From these genes, we identified 89 possible molecular chaperones that matched known molecular chaperones from other organisms available in various databases such as Uniprot, Gene Ontology, cpnDB and NCBI. For an in-depth analysis of molecular functions, we used homologous clustering to transfer knowledge from unknown to known functions using Cluster Analysis of Sequences (CLANS) bioinformatics software. The results reveal 12 major groups of chaperones that contribute to the cold-adaptation mechanism through their molecular function, biological processes and cellular components. The findings lay the foundation for future functional genomics studies on this organism and shed light on how lower eukaryotic cells respond to low temperature.

Key words: Molecular chaperones, clustering analysis, functional genomics, *Glaciozyma antarctica*

INTRODUCTION

Glaciozyma antarctica is a psychrophilic yeast that is able to grow and proliferate in extremely cold environments (Hashim *et al*., 2013). To survive in such environments, this yeast must be able to protect its proteins from improper folding, stabilise the fluidity of its membrane, produce low-temperature active proteins and maintain the negative supercooling level of its DNA structure. Molecular chaperones are one of the most important proteins that enable *G. antarctica* to survive in extreme environments. These proteins are known as folding modulators that play a vital role in the conformational quality control of the proteome by interacting with, stabilising and regulating protein conformational states (Buchner, 1996). Chaperones capture unfolded polypeptides, stabilise them, and prevent misfolding from accumulating in stressed cells. Many chaperones are upregulated upon heat shock or cold shock or in response to other insults that promote protein misfolding (Hartl *et al*., 2011). It is believed that chaperones provide the key survival mechanism of extremophiles at low temperature, such as in the Arctic and Antarctic or during cold winters that cause harmful effects through extreme temperature drops (Relina and Gulevsky, 2003). Therefore, it is important to investigate the mechanism that is used by *G. antarctica* to overcome temperature downshifts and maintain their protein activities without being interrupted by cellular stress.

The Whole Genome Sequencing (WGS) project has identified a total of 7857 genes from *G. antarctica* (http://www.genomemalaysia.gov.my/glaciozyma_antarctica/). In genomic studies, data collection is the bottleneck as researchers are bombarded with large quantities of genomic data. Tremendous amounts of genome data can occasionally lead to fruitless research if no realistic approach is adopted. Indeed, identifying functionally relevant genes to study presents a major challenge. We have overcome this bottleneck from data collection to analysis by large-scale genomic data mining, using many potentially diverse datasets from public repositories to address a specific biological question. In this study, we performed a bioinformatics analysis to identify

* To whom correspondence should be addressed.

possible molecular chaperones out of thousands of genes from our genome database in a reasonable amount of time and transfer the knowledge of protein structure and function using a homologous relationship with known proteins. The three main objectives of this study were as follows: 1) to identify all possible molecular chaperones by developing an effectual method for bioinformatics analysis, 2) to transfer the knowledge of sequence to functions by gathering data from diverse available repositories, and 3) to analyse the function of possible molecular chaperones in cold adaptation and stress response. We believe that this new dataset will provide a useful resource for future genetic and genomic studies and expression analysis of *G. antarctica*.

MATERIALS AND METHODS

Bioinformatics-based screening and selection of target proteins from *G. antarctica*

The genome of *G. antarctica* (http://www.genomemalaysia.gov.my/glaciozyma_antarctica/) contains 7857 genes, with at least 10% being novel or exhibiting no detectable sequence similarity to known folds. Data mining of all molecular chaperones registered in various available databases, for example, Uniprot, Gene Ontology, cpnDB and NCBI, was performed. Key words such as 'chaperone', 'chaperonin', 'heat shock protein' (hsp) and 'cold shock protein' (csp) were used in the search process. Subsequently, all possible sequences of molecular chaperones were used to search for *G. antarctica* molecular chaperones by running BLAST searches within the *G. antarctica* genome to find similar sequences using Unix commands, Perl Script, PuTTY and WinSP. Aspects that were considered were the query name, % identity, e-value, query length and match length. A cutoff *e*-value of 10^{-5} or less as set in the script. Filtering was performed by establishing that the percentage length (match length/query length) should be more than 50%.

Clustering for homologs using CLANS map

Molecular chaperone sequences were retrieved from GenBank (http://www.ncbi.nlm.nih.gov) by running BLAST searches using reference molecular chaperones sequences from *Saccharomyces cerevisiae* and *Cryptococcus neoformans*. The molecular chaperones dataset included a total of 234 sequences. Cluster Analysis of Sequences (CLANS) was used to identify subfamilies of closely related molecular chaperone sequences and elucidate the relationships between and within the amino acid pairwise sequence similarities with the closest families, which were *S. cerevisiae* and *C. neoformans* molecular chaperones. CLANS is a Java utility based on the Fruchterman-Reingold graph layout algorithm. It runs BLAST searches on given sequences in an all-against-all fashion and clusters them in 3D according to their similarity.

Functional analysis

The assembled sequences were compared against the NCBI nr and nt database and the Swiss-Prot database using BLASTN with an *e*-value $< 10^{-5}$. To annotate the assembled sequences with GO terms, the Swiss-Prot BLAST results were imported into BLAST2GO, allowing gene functions to be determined and compared. Interesting sequences were identified by the BLAST results against the database with a cutoff *e*-value of $<10^{-5}$. Genes from other yeasts such as *S. cerevisiae and C. neoformans* were used as references.

RESULTS AND DISCUSSION

The genes were annotated by aligning with the deposited ones in diverse protein databases, including the National Center for Biotechnology Information (NCBI) non-redundant protein (nr), the NCBI non-redundant nucleotide sequence (nt), the UniProt/Swiss-Prot and the Gene Ontology (GO) databases using BLASTX with a cutoff *e*-value of 10^{-5}. In total, 89 genes were successfully identified as molecular chaperones or heat-shock proteins. Table 1 lists the *G. antarctica* molecular chaperones identified in this study.

CLAN clustering analysis was performed to identify the functions of the possible molecular chaperones (Fig. 1). The clustering was performed using pairwise sequence similarities with known molecular chaperones from *S. cerevisiae* and *C. neoformans*. There were 12 main groups that had a P-value near 0.

The group that was clustered with the best P-value was a mega-complex chaperonin well known as T-complex protein-1 ring chaperonin (TRiC). This complex consists of two identical stacked rings, each containing eight different proteins. Unfolded polypeptides enter the central cavity of the complex and are folded in an ATP-dependent manner. The complex folds various proteins, including actin and tubulin (Brackley and Grantham, 2009). Pucciarelli *et al.* (2006) found that the folding mechanism of Antarctic fish TRiC differed from that of mammalian TRiC. Their studies described the necessity of structural flexibility for catalytic activity and the concomitant hazard of cold-induced denaturation. The second group identified in the clustering analysis was the HSP70 group. HSP70 proteins, a family of conserved ubiquitously expressed heat-shock proteins, are the central components of the cellular network of molecular chaperones and

Table 1. List of *G. antarctica* molecular chaperones

Molecular Chaperones	*G. antarctica* gene identification	Protein Description	Molecular Chaperones	*G. antarctica* gene identification	Protein Description
TCP-1 chaperonin	LAN_03_156	TCP1 delta subunit	J-containing domain proteins	LAN_03_146	J domain-containing protein
	LAN_03_211	TCP1 alpha subunit		LAN_03_253	J domain-containing protein
	LAN_10_409	Cpn60		LAN_03_415	J domain-containing protein
	LAN_11_354	FYVE-type zinc finger domain-containing protein/Cpn60		LAN_03_528	J domain-containing protein
				LAN_03_565	J domain-containing protein
				LAN_04_115	J domain-containing protein
	LAN_11_451	TCP-1 eta subunit		LAN_04_375	J domain-containing protein
	LAN_11_508	TCP-1 beta subunit		LAN_05_107	J domain-containing protein
	LAN_12_031	TCP-1 gamma subunit		LAN_05_286	J domain-containing protein
	LAN_12_520	TCP-1 epsilon subunit		LAN_05_287	J domain-containing protein
	LAN_16_575	TCP-1 theta subunit		LAN_05_432	J domain-containing protein
	LAN_16_879	TCP-1 zeta subunit		LAN_05_464	J domain-containing protein
				LAN_05_545	CR-type zinc finger and J domain-containing protein
HSF proteins	LAN_02_254	Heat-shock protein			
	LAN_03_001	HSF family protein		LAN_06_350	J domain-containing protein
	LAN_03_640	HSF family protein		LAN_08_053	J domain-containing protein
	LAN_03_643	HSF family protein		LAN_08_145	J domain-containing protein
	LAN_03_655	HSF family protein		LAN_09_046	Protein psi1 homolog with DnaJ domain
	LAN_03_716	HSF family protein			
	LAN_03_733	HSF family protein		LAN_10_093	hscB family protein
	LAN_04_495	HSF family protein		LAN_10_106	J domain-containing protein
	LAN_06_169	HSF family protein		LAN_11_316	DnaJ and SEC63 domain-containing protein
	LAN_08_425	HSF family protein			
	LAN_09_178	HSF family protein		LAN_11_507	Heat-shock protein
	LAN_14_288	HSF family protein		LAN_13_055	DPH4 family protein
	LAN_16_088	HSF family protein		LAN_13_365	J domain-containing protein
	LAN_16_580	HSF family protein		LAN_16_061	J domain-containing protein
	LAN_17_157	Heat-shock protein		LAN_16_170	J domain-containing protein
				LAN_16_195	J domain-containing protein
HSP70 proteins	LAN_10_353	Heat-shock protein		LAN_16_519	J domain-containing protein
	LAN_01_088	Heat-shock protein		LAN_17_136	J domain-containing protein
	LAN_09_065	Heat-shock protein		LAN_09_124	J domain-containing protein
	LAN_10_316	Heat-shock protein		LAN_12_227	J domain-containing protein
	LAN_12_055	Uncharacterised protein			
	LAN_13_463	Heat-shock protein			
	LAN_15_171	Heat-shock protein			
	LAN_16_202	Heat-shock protein			
CS-domain-HSP20-like chaperone that binds HSP90	LAN_12_053	CS-domain containing protein	Cold-shock proteins	LAN_16_627	Cold-shock domain-containing protein
				LAN_16_676	CSD (cold-shock) domain-containing protein
				LAN_16_853	CSD (cold-shock) domain-containing protein
HSP90 proteins	LAN_16_646	Endoplasmin heat-shock protein	AAA proteins	LAN_14_065	AAA ATPase
				LAN_16_532	Chaperonin clpA/B with AAA ATPase
	LAN_10_287	Heat-shock protein			
				LAN_01_186	clpA/clpB family protein
CS-domain proteins	LAN_11_064	CS domain-containing protein	Ubiquitin proteins	LAN_12_331	UBA and ubiquitin-like domain-containing protein
	LAN_16_840	CS domain-containing protein			
TPR domain proteins	LAN_10_408	Tetratricopeptide repeat-containing protein			

folding catalysts that exist in all living organisms. HSP70s are a crucial part of the cell's machinery for protein folding that helps to protect cells from stress. ATP binding and hydrolysis are essential for HSP70 protein activity, where the ATPase cycle is controlled by co-chaperones of the family of J-domain proteins (Mayer and Bukau, 2005). Studies show that HSP70 chaperone performs many diverse roles in the cell, including folding of nascent proteins, translocation of polypeptides across organelle membranes, coordinating responses to stress, and targeting selected proteins for degradation. Our clustering analysis revealed a close relationship between HSP70 families and

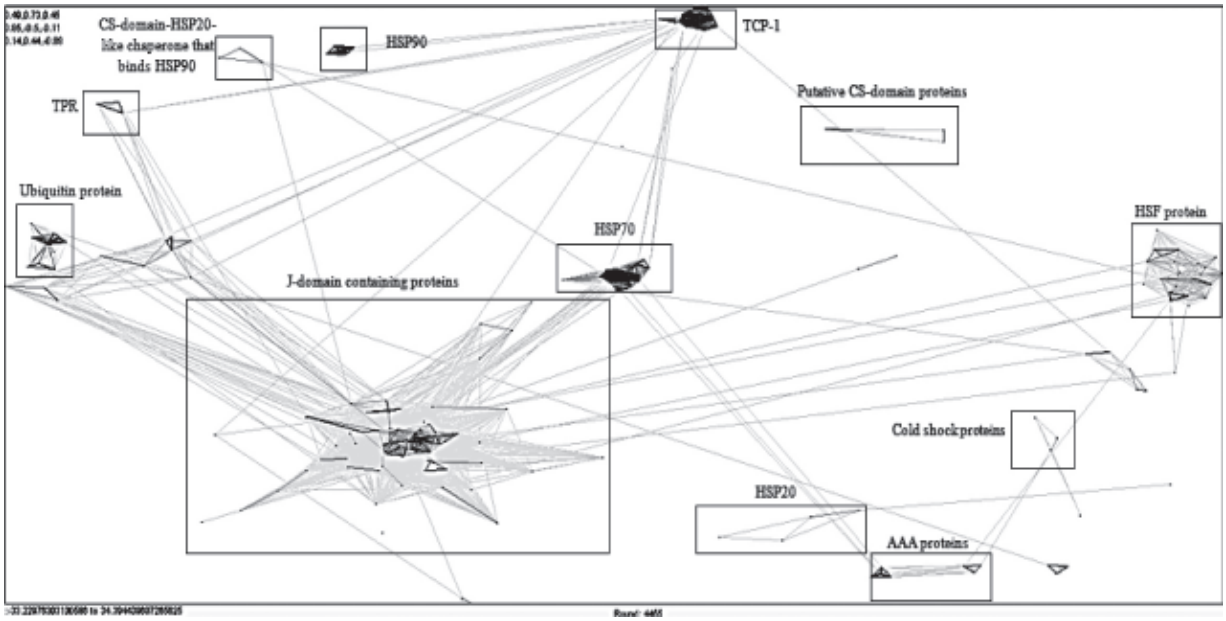

Fig. 1. Clustering analysis of all possible molecular chaperones of *G. antarctica* using molecular chaperones from *S. cerevisiae* and *C. neoformans* as references. All molecular chaperones, including 89 possible molecular chaperones from *G. antarctica*, 72 molecular chaperones from *C. neoformans* and 73 molecular chaperones from *S. cerevisiae*, were clustered based on sequence similarities. The known functions of molecular chaperones from the references serve as guidelines for identifying the functions of the *G. antarctica* molecular chaperones.

J-domain proteins. The J-domain proteins were clustered as the largest group of molecular chaperones in *G. antarctica*. The J-domain protein is a member of the HSP40 family of molecular chaperones, which is also called DnaJ, the members of which regulate the activity of HSP70s. DnaJ or HSP40 binds to DnaK of HSP70 and stimulates its ATPase activity, generating the ADP-bound state of DnaK, which interacts stably with the polypeptide substrate (Greene *et al*., 1997). The third identified group of molecular chaperones was HSP90, which is closely related to a co-chaperone that contains CS-domain-HSP20-like chaperones and a TPR domain. HSP90 is a chaperone protein that assists other proteins in folding properly, stabilises proteins against heat stress, and aids in protein degradation. HSP90 is one of the most common heat-related proteins, which are the most highly expressed cellular proteins across all species that protect cells when stressed by elevated temperatures (Pearl and Prodromou, 2006). The fourth clustered group of molecular chaperones was the ubiquitin-like-binding proteins that assist in cell cycling, responding to DNA replication stress and regulating the degradation of proteins via proteolysis. The fifth major group was heat-shock factors (HSFs). HSFs are transcriptional activators of heat-shock genes in eukaryotes. In the absence of cellular stress, HSF is inhibited by association with heat-shock proteins and therefore not active. Cellular stresses such as increased temperature cause proteins in the cell to misfold. Heat-shock proteins bind to the misfolded proteins and dissociate from HSF (Parsell and Lindquist, 1993). The other small groups of molecular chaperones identified were the HSP20 and ATPases associated proteins. In *Sulfolobus solfataricus* P2, the HSP20 are ubiquitious chaperones that promote thermotolerance by enhancing protein synthesis in *E. coli*, which features HSPs. These proteins protect *E. coli* proteins from heat denaturation through their chaperone activity (Li *et al.*, 2012).

Clustering analysis also demonstrated that the presence of small HSPs (sHSPs) reflects the response mechanism of organisms to some extreme stresses existing in the environment. sHSPs have been suggested to contribute to thermal resistance. In the present study, 18 genes were found to have similarities to HSPs of *S. cerevisiae* and *C. neoformans*. Among them, 'protein folding' and 'ATP hydrolysis' were primarily involved in the identified HSPs. Importantly, 'response to stress' was detected in the HSPs, which suggests that HSPs may be involved in cellular stress resistance in *G. antarctica*, act as molecular chaperones that block the aggregation of unfolded proteins and have a cytoprotective function under stressful situations.

The most interesting and central finding of our study was the presence of cold-shock proteins in *G. antarctica*. Cold-shock proteins are postulated to help the cell to survive at temperatures below their optimum growth temperature, in contrast to heat-

shock proteins, which help the cell to survive at temperatures above the optimum. Studies show that cold shock affects membrane composition and organisation to maintain the optimum membrane function (Phadtare *et al.*, 1999). Cold-induced gene expression has been observed in eukaryotic organisms, such as the cold-shock proteins identified in yeast and the antifreeze proteins in Antarctic fish (Thieringer *et al*., 1998).

In conclusion, this work has identified 89 possible molecular chaperones from *G. antarctica* genome and grouped them into different classes based on their sequences' similarity to those of other molecular chaperones. This data will facilitate research on molecular chaperones in *G. antarctica*, especially on their roles and regulation in protecting cells from drastic temperature fluctuations.

ACKNOWLEDGEMENTS

We thank the Malaysian Genome Institute (MGI) for providing access to the *Glaciozyma antarctica* Genome Database. The authors would like to acknowledge the financial support from the Ministry of Science and Technology, Malaysia (MOSTI) through grants 02-05-20-SF0007 and 10-05-16-MB002.

REFERENCES

Brackley, K.I. & Grantham, J. 2009. Activities of the chaperonin containing TCP-1 (CCT): implications for cell cycle progression and cytoskeletal organisation. *Cell Stress and Chaperones,* **14**: 23-31.

Buchner, J. 1996. Supervising the fold: functional principle of molecular chaperones. *FASEB Journal*, **10**: 10-19.

Greene, M.K., Maskos, K. & Landry, S.J. 1997. Role of the J-domain in the cooperation of Hsp40 with Hsp70. *PNAS*, **95(11)**: 6108-6113.

Hartl, F.U., Bracher, A. & Hayer-Hartl, M. 2011. Molecular chaperones in protein folding and proteostasis. *Nature*, **475**: 324-332.

Hashim, N.H.F., Bharudin, I., Law, D.S.N., Higa, S., Bakar, F.D.A., Nathan, S., Rabu, A., Kawahara, H., Illias, R.M., Najimudin, N., Mahadi, N.M. & Murad, A.M.A. 2013. Characterization of Afp1, an antifreeze protein from the psychrophilic yeast *Glaciozyma antarctica* PI12. *Extremophiles*, **17**: 63-73.

Li., D.C., Yang, F., Lu, B., Chen, D.F. & Yang, W.J. 2012. Thermotolerance and molecular chaperone function of the small heat shock protein HSP20 from hyperthermophilic archaeon, *Sulfolobus solfataricus* P2. *Cell Stress and Chaperones*, **17**: 103-108.

Mayer, M.P. & Bukau, B. 2005. Hsp70 chaperones: cellular functions and molecular mechanism. *Cellular and Molecular Life Sciences*, **62(6)**: 670-684.

Parsell, D.A. & Lindquist, S. 1993. The function of heat-shock proteins in stress tolerance: degradation and reactivation of damaged proteins. *Annual Review of Genetics*, **27**: 437-496.

Pearl, L.H. & Prodromou, C. 2006. Structure and mechanism of the Hsp90 molecular chaperone machinery. *Annual Review of Biochemistry*, **75**: 271-294.

Phadtare, S., Alsina, J. & Inouye, M. 1999. Cold-shock response and cold-shock proteins. *Current Opinion in Microbiology*, **2(2)**: 175-180.

Pucciarelli, S., Parker, S.K., Detrich, H.W. & Melki, R. 2006. Characterization of the cytoplasmic chaperonin containing TCP-1 from the Antarctic fish *Notothenia coriiceps*. *Extremophiles*, **10**: 537-549.

Relina, L.I. & Gulevsky, A.K. 2003. A possible role of molecular chaperones in cold adaptation. *Cryoletters*, **24(10)**: 203-212.

Thieringer, H.A., Jones, P.G. & Inouye, M. 1998. Cold shock and adaptation. *BioEssays*, **20**: 49-57.